콘텐츠 중심 네트워킹을 적용한 무선 센서 네트워크 환경의 미세먼지 통계 예보 시스템

전유정, 정재훈
성균관대학교 소프트웨어학과
{dbwjd3508, pauljeong}@skku.edu*

PM10 statistical forecasting system in CCN-applied WSN

YuJeong Jeon and Jaehoon (Paul) Jeong*
Department of Software Sungkyunkwan Univ.

요 약

본 논문은 CCN(Content Centric Network) 환경에서 SVM(Support Vector Machine) 모델 기반의 미세먼지 통계 예보 시스템을 소개한다. 현재 기상청에서 미세먼지(PM10) 등급을 예보하는 과정 중 농도에 영향을 미치는 기상 및 대기질 측정값들을 각각 회귀모델, 신경망모델, 의사결정모델에 적용하여 예보 농도를 산출하는 통계 예보 모델은 기존의 TCP/IP방식을 따르는 전형적인 네트워크 과정을 거쳐 데이터를 수집하여 진행한다. 하지만 현재 미세먼지 예보센터에서 사용하는 슈퍼컴퓨터는 기상 예보에 사용되는 것보다 성능이 1/100배가량 떨어져 처리 속도 면에서 비효율적이다. 따라서 지역별 각 서버의 과부하를 막고 라우터에서 동일한 데이터를 요구하는 패킷의 중복된 전송을 막는 CCN을 활용하여 데이터를 수집하는 속도를 늘리고, 현재 이용되는 통계 예보 모델이 아닌 시계열 데이터에서 보편적으로 사용하는 ARIMA와 LSTM, 비선형 SVM 모델을 사용하여 미세먼지 예보 과정의 성능과 정확성을 높인다.

Ⅰ. 서론

동아일보가 공개한 기상청의 미세먼지 예보센터에서 예보 과정을 촬영한 영상에 따르면 미세먼지 예보 정확도가 80%이상인 선진국에 비해 우리나라는 70%대에만 머무를 수밖에 없는 이유는 다음과 같다[1]. 현재 미세먼지 예보 모델링에 쓰이는 슈퍼컴퓨터는 기상청에서 쓰이는 슈퍼컴퓨터의 성능의 1/100에 미치지 못하고, 모델링을 시작하게 되면 하루 네 번 진행되는 3시간 동안 데이터의 변경 사항을 반영을 하지 못한다. 또한 미세먼지를 유발하는 주요 장소인 전국 600여개의 사업장의 데이터를 반영하지 못해 예보관이 모델링 결과에 이를 별도로 고려해 결정한다. 이러한 과정에서 예보의 정확성은 떨어질 수밖에 없다. 따라서 본 논문에서는 위에서 언급된 문제 원인 중 현재 쓰이는 데이터뿐만 아니라 미세먼지가 발생하는 주요 사업장의 측정값을 포함한 새로운 통계 예보 모델을 구현하고, 차세대 네트워크로 활발히 연구되고 있는 CCN 토폴로지[2]를 활용하여 관측소별 환경 데이터를 주고받는 환경을 제안하고자 한다.

국립환경과학원은 미국 환경청(EPA)이 개발한 모델을 바탕으로 하여 통계예보모델을 회귀 모델(Regression), 신경망 모델(Neural network), 의사 결정법 모델(Decision tree)을 개발하였다[3]. 서울시에서는 이 통계예보모델을 기반으로 “서울특별시 먼지 예보 및 경보에 관한 조례(서울특별시조례 제4247호)”를 제정했다. 현재 구축된 미세먼지 예보모델은 통계예보 모델에 속한 모델을 모두 수행한 뒤, 그 결과값과 화학 수송 예보 모델값을 함께 활용하여 2차 회귀모델을 거쳐 각각 가중치를 두어 최종 예보 결과 값을 도출한다. 통계 예보 모델에서 쓰이는 세 모델에 대한 설명은 다음과 같다[4].

- 의사결정모델: 유사한 오염농도를 가지는 군집을 자동적으로 분류
- 회귀분석모델: 기상 및 대기질 측정값을 변수로 사용해 과거의 추이에 대한 방정식을 만들어 예측하는 방법
- 인공신경망모델: 기상측정값 및 대기질 측정농도를 입력값으로 사용해 비선형 방정식을 통해 입력인자에 가중치를 두어 예측하는 방식으로 현재 유숙현(2011)이 제안한 설계를 따라 입력층, 은닉층1, 2, 출력층의 네 개 층으로 구성되어 있으며, 출력층을 제외한 각층의 노드 수는 입력 인자들에 따라 다양하게 변하도록 하였다[4].

본 논문은 미세먼지 영향 인자로 보고되는 기상데이터를 기반으로 1시간 단위의 기온, 강우, 상대 습도, 풍속, 풍향 기상 데이터와 과거 농도 데이터를 이용해 LSTM, ARIMA, SVM로 분석한 후, 예측하는 데 걸린 시간과 정확도를 비교한다. 정확도는 2018년 3월부터 개정된 미세먼지 예보 기준을 적용해 좋음, 보통, 나쁨, 매우 나쁨의 4단계로 분류하여 정확도를 산출한다. 더불어 PARC(Palo Alto Research Center)에서 오픈소스로 제공되는 CCNx library[5]를 이용하여 지역별 가상의 서버를 설정하고 클라이언트에서 각 서버에서 다량의 미세먼지 데이터를 요청하고 수집하여 각 모델링을 수행하는데 이용하고 센서 네트워크 환경에서 속도와 안정성 면에서 CCN의 성능을 검증한다.

Ⅱ. 본론

1. 통계 예보 모델 설계

미세먼지 농도는 시간에 따른 데이터에 가중치가 결정된다. 일반적으로 가장 가까운 과거 시간의 데이터가 가중치가 높고, 오래된 데이터일수록 상대적으로 예측값에 영향을 덜 끼친다[6]. 연속성과 상하 편이를 가진 시계열 자료를 기반으로 하는 분석이므로 시계열 데이터를 다루는 전형적인 모델인 RNN (Recurrent Neural Network)[7]와 회귀와 분석에 이용되는 SVM (Support Vector Machine)[8]을 비선형 데이터인 시계열 자료를 학습할 수 있는 모델로 개선한 시스템을 통계예보모델로 적용한다. 3년 간 시간별 미세먼지 데이터 셋과 기상 데이터를 학습해 분석을 진행한 후 세 모델의 예측력, 속도 면에서 비교하는 동시에 기존에 사용되고 있는 모델들에서 분석했던 변수들 간의 상관관계에서 새로운 모델을 도입했을 때 어떤 변화가 있는지 RMSE로 파악한다. 파악한 내용을 바탕으로 정확도를 개선하기 위한 모델링 시스템을 고안하여 이를 적용한다. 향후 미세먼지 농도를 예측하기 위해서 실시간 미세먼지 데이터와 기상 데이터, 주요 사업장 관측값을 수집해 분석에 사용할 수 있도록 구조화한다. 각 모델의 설계와 성능에 영향을 미치는 주요 변수 설정은 다음과 같다.

① LSTM

- 설계 방식: 정규화된 데이터를 Sequantial LSTM과 Supervised learning을 적용한 LSTM을 구현, 실험하여 둘 중 최적의 RMSE와 정확도를 가지는 모델의 값을 기록한다. Giovanni(2007)에 따르면, hidden layer의 개수가 많을수록, epoch가 클수록, variate가 많을수록, 즉 Multivariate LSTM의 성능이 좋다[9].
- 변수 설정: epoch(100), hidden layer의 개수(100), batch size(24)

② ARIMA

- 설계 방식: 데이터를 안정적으로 가공하기 위해 Dicky-Fuller test와 이동 평균(rolling mean), 이동 표준편차(rolling standard deviation) 산출 과정을 거친다. 이후 데이터를 안정적으로 만들기 위해 데이터에 log와 차분을 적용한 값을 최종 훈련 데이터로 활용한다. ACF, PACF 함수를 구하여 가장 적합한 모듈은 ARIMA이며 적절한 p, d, q값이 1, 1, 1임을 파악한다.

③ SVM

- 설계 방식: SVM에 적용하기 위해 데이터를 Sliding window method로 가공하고, Standardscalar로 정규화한 데이터를 가우시안 커널 함수를 적용한 SVM 모델에 적용한다. 이후 시계열 데이터 형식에 맞게 끔 데이터를 재형성한다.
- 변수 설정: C(100), gamma(0.01)

2. CCN 평가 시나리오 설정 및 실제 적용

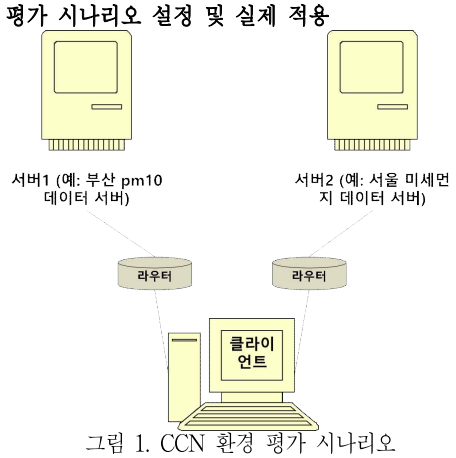

그림 1. CCN 환경 평가 시나리오

그림 1과 같이 기기 3대를 이용하여 2대는 서버로 미세먼지 데이터를 보유하고, 1대는 클라이언트로 데이터를 요청, 수집하는 담당을 맡는다. 본 논문에서는 CCN Distillery 라이브러리를 이용하며, CCN Forwarding engine으로는 통합 라이브러리인 CCNx-Distillery에 포함된 Athena forwarder를 이용하여 리눅스 커널 상에서 라우팅한다. 검증 후 실제 적용을 위해 모델링 결과를 반영한 예보 등급을 나타내는 프로그램을 파이썬 기반 웹 프레임워크인 Django framework를 이용하여 구축한다. 또한 최근의 미세먼지 측정값을 제공하는 에어코리아에서 실시간으로 제공하는 미세먼지 농도 데이터와 기상 데이터를 selenium 모듈을 활용해 크롤링(crawling)하여 추출, 가공하여 모델의 훈련 데이터로 사용한다.

Ⅲ. 결론

모델 / 크기	LSTM		ARIMA		SVM	
	RMSE	ACC	RMSE	ACC	RMSE	ACC
15일	3.431	81.3	0.479	93.7	3.817	88.7
1개월	5.300	77.1	0.186	97.7	8.580	98.6
3개월	7.529	89.0	10.997	92.9	8.582	89.8
6개월	4.911	93.3	4.639	93.8	8.986	90.3
1년	9.117	95.9	22.396	82.4	7.208	82.9
3년	9.678	92.4	40.292	78.0	8.836	85.4

표 1. 데이터 셋의 크기별 각 모델의 RMSE, 정확도

표1은 데이터의 크기에 따른 각 모델의 RMSE와 정확도를 나타낸다. 각 모델의 RMSE와 정확도는 모든 데이터 셋을 적용할 시, 변화율이 큰 데이터와 작은 데이터를 번갈아 사용한 결과값의 평균을 산출한 것이다. 그 결과, 세 모델 모두 데이터의 크기가 증가할수록, 데이터의 변화가 급격할수록 RMSE는 커지고 정확도는 다소 떨어지는 것을 보였다. 특히 ARIMA의 경우, RMSE가 급격하게 커지면서 초반의 가장 높았던 정확도를 가진 것에 비해 후반에는 가장 낮은 정확도와 높은 RMSE를 보이며 성능이 떨어졌다. 반면 LSTM은 데이터 셋이 작거나 변화폭이 클수록 다소 정확도가 떨어졌지만 데이터 셋이 커질수록 가장 정확도가 높았다. 또한 LSTM과 ARIMA, SVM의 실행 시간은 6개월 데이터를 기준으로 각각 162초, 4.01초, 2.06초가 소요되어 모든 데이터 셋에서 LSTM이 가장 느렸다. 결론적으로, LSTM은 실시간으로 미세먼지 데이터를 모델링해서 예보하는 데에는 속도와 성능 면에서 두 모델에 비해 적합하지 않다. 반면 ARIMA는 적은 양의 데이터에는 우수한 예측력을 가졌지만 데이터 셋이 커질수록 예측값과 실측값의 오차가 커져 정확도가 떨어진다. 따라서 빠른 계산 속도와 데이터 셋의 변화추이와 크기에 영향을 적게 받는 SVM 모델이 적합하다는 결론이 도출된다.

오늘날 시행되고 있는 미세먼지 예보 모델은 성능이 떨어지는 기기를 쓸 뿐만 아니라 인적 자원 및 기술력의 부족, 모델링에 반영되지 못한 데이터 등 다각도에서 정확도를 떨어뜨린다는 지적이 나오고 있다. 본 논문에서는 예보 정확도를 떨어뜨리는 원인 중 통계예보모델의 떨어지는 정확성과 성능을 개선하기 위해 과거의 시계열 데이터에서 그 변화의 추이를 예상하고자 전형적으로 사용되고 있는 LSTM, ARIMA과 함께 가우시안 커널을 적용한 SVM 모델링을 적용하는 동시에 반영되지 못했던 주요 변수를 반영하는 신모델을 제안한다. 더불어 CCN 환경 설계를 통해 향후 거대해질 기상 관측 네트워크에서 서버에 저장되었거나 센서로부터 들어오는 데이터가 단 한 번의 요청으로 모델링에 필요한 데이터가 예보 센터에 효과적으로 도달할 수 있음을 검증하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학 지원사업의 연구결과로 수행되었음(2015-0-00914).

참 고 문 헌

[1] 미세먼지 예보, 3명이 토의해 결정, <동아일보>, 2018-04-26 03:00, http://news.donga.com/3/all/20180426/89798548/1#csidxdda09fbab8515bc9247979e4f5db750(2018-05-07 13:00)

[2] Cicn(2018), Retrieved May, 24, 2018, https://wiki.fd.io/view/Cicn

[3] Mike Abraczinskas et al, Guidelines for Developing an Air Quality (Ozone and PM2.5) Forecasting Program(North Carolina, U.S. EPA Office of Air Quality Planning and Standards Information Transfer and Program Integration Division AIRNow Program Research Triangle Park, 2003)

[4] 구윤서 외 미세먼지 예보시스템 개발, 한국대기환경학회지 제26권 제6호, (2010)

[5] CCNx_Distillery(2016), Retrieved May, 24, 2018, https://github.com/PARC/CCNx_Distillery

[6] 채희정, 미세먼지예보 정확도 향상에 관한 연구, 2011, 조선대학교 공학박사학위논문

[7] Hardik Goel et al, Residual Recurrent Neural Networks for Multivariate Time Series Forecasting, University of Minnesota (2017)

[8] Steinwart et al, Support Vector Machines, Springer, 2008

[9] Giovanni Raimondo et al, A MACHINE LEARNING TOOL TO FORECAST PM10 LEVEL, Polytechnic of Turin, Italy and Earth Science Centre of Gothenburg, Sweden(2007)